Cost-Effective Discovery of Nucleotide Polymorphisms in Populations of an Allopolyploid Species Using Pool-Seq

Hirao, Akira S ; Onda, Yoshihiko ; Shimizu-Inatsugi, Rie ; Sese, Jun ; Shimizu, Kentaro K ; Kenta, Tanaka

Abstract: Population genetics studies of allopolyploid species lag behind those of diploid species because of practical difficulties in analysis of homeologs-duplicated gene copies originating from hybridized parental species. Pool-Seq, i.e. massive parallel sequencing of pooled individuals, has high potential for detecting nucleotide polymorphisms within and among multiple populations; however, its use has been limited to diploid species. We applied Pool-Seq to an allopolyploid species by developing a bioinformatic pipeline that assigns reads to each homeolog as well as to each polymorphic allele within each homeolog. We simultaneously sequenced eight genes from twenty individuals from each of 24 populations, and found over 100 polymorphic sites in each homeolog. For two sites, we estimated allele frequencies using the number of reads and then validated these estimations by making individual-based estimations. Pool-Seq using our bioinformatic pipeline allows efficient evaluation of nucleotide polymorphisms in a large number of individuals, even in allopolyploid species.

DOI: https://doi.org/10.4236/ajmb.2017.74012

Scientific Research Publishing

American Journal of Molecular Biology, 2017, 7, 1031-1046
http://www.scirp.org/journal/ajmb
ISSN Online: 2161-6663
ISSN Print: 2161-6620

Cost-Effective Discovery of Nucleotide Polymorphisms in Populations of an Allopolyploid Species Using Pool-Seq

Akira S. Hirao[1*], Yoshihiko Onda[2,3,4], Rie Shimizu-Inatsugi[5], Jun Sese[6], Kentaro K. Shimizu[4,5], Tanaka Kenta[1]

[1]Sugadaira Research Station, Mountain Science Center, University of Tsukuba, Ueda, Japan
[2]Sugadaira Montane Research Center, University of Tsukuba, Ueda, Japan
[3]Cellulose Production Research Team, Biomass Engineering Research Division, RIKEN Center for Sustainable Resource Science, Yokohama, Japan
[4]Kihara Institute for Biological Research, Yokohama City University, Yokohama, Japan
[5]Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland
[6]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan
Email: *akihirao@sugadaira.tsukuba.ac.jp

How to cite this paper: Hirao, A.S., Onda, Y., Shimizu-Inatsugi, R., Sese, J., Shimizu, K.K. and Kenta, T. (2017) Cost-Effective Discovery of Nucleotide Polymorphisms in Populations of an Allopolyploid Species Using Pool-Seq. *American Journal of Molecular Biology*, **7**, 1031-1046.
https://doi.org/10.4236/ajmb.2017.74012

Received: April 11, 2017
Accepted: September 11, 2017
Published: September 14, 2017

Abstract

Population genetics studies of allopolyploid species lag behind those of diploid species because of practical difficulties in analysis of homeologs-duplicated gene copies originating from hybridized parental species. Pool-Seq, *i.e.* massive parallel sequencing of pooled individuals, has high potential for detecting nucleotide polymorphisms within and among multiple populations; however, its use has been limited to diploid species. We applied Pool-Seq to an allopolyploid species by developing a bioinformatic pipeline that assigns reads to each homeolog as well as to each polymorphic allele within each homeolog. We simultaneously sequenced eight genes from twenty individuals from each of 24 populations, and found over 100 polymorphic sites in each homeolog. For two sites, we estimated allele frequencies using the number of reads and then validated these estimations by making individual-based estimations. Pool-Seq using our bioinformatic pipeline allows efficient evaluation of nucleotide polymorphisms in a large number of individuals, even in allopolyploid species.

Keywords

Arabidopsis kamchatica, Allele Frequency, Homeolog, Individual-Based Genotyping, Massive Parallel Sequencing

1. Introduction

Allopolyploid species, which result from hybridization of closely related taxa accompanied with whole-genome duplication, are rare in amniotes but frequent in fish and amphibians and common in plants [1] [2]. Particularly in plants, allopolyploid species are considered to have greater ecological adaptability to broader and novel environmental niches [3] [4]. Many high-yield crops are allopolyploid species, including bread wheat (*Triticum aestivum* L.), canola (*Brassica napus* L.), cotton (*Gossypium hirsutum* L.), tobacco (*Nicotiana tabacum* L.), and coffee beans (*Coffea arabica*). The genetic basis of adaptation in wild and cultivated allopolyploid species has received broad attention in various fields such as evolutionary biology, agriculture, and biotechnology. However, genetic studies of allopolyploid species lag behind those of diploids because of difficulties in analyzing homeologs-duplicated gene copies originating from the two parental species. Practical difficulties in sequencing homeologs limit the development of genetic resources in allopolyploid species. Indeed, the draft genome (ordered along the chromosomes) of the allopolyploid crop with the most economic value, *T. aestivum* [5], was released more than 10 years after the draft genome of the model plant *Arabidopsis thaliana*.

Since the 2010s, however, the shortcoming in genetic resources for allopolyploids has changed because of rapid advances in sequencing and genomics technologies [6]. Massive parallel sequencing and computational discrimination of homeologs, without homeolog-specific PCR procedures, has started to enable genome-wide studies in allopolyploid organisms, such as whole genome sequencing [7], gene expression profiling [8] [9], and genome evolution [10]. Despite such genome-wide intra- and/or inter-individual data becoming available, obtaining intra- and/or inter-population genetic variation in allopolyploids cost-effectively for population genetics studies remains a challenge [11]. One efficient solution to obtain population-level data is to sequence pools of individuals, namely Pool-Seq [12]. However, its application has been limited to diploid species e.g. [13] [14], because of difficulties in applying a bioinformatic pipeline for allopolyploid species that simultaneously discriminates homeolog-specific polymorphisms (defined as base differences between the subgenomes within a polyploid genome [15]) as well as allele-specific polymorphisms in each homeolog within/among populations. In this report, we developed a Pool-Seq protocol to apply to natural populations of an allopolyploid species, *Arabidopsis kamchatica* (DC.) K. Shimizu et Kudoh subsp. *kamchatica*.

Arabidopsis kamchatica—a wild relative of the model plant *A. thaliana*—is an allotetraploid ($2n = 4x = 32$) species with a broad habitat range, wide altitudinal distribution [16], lowland seaside and lakeside sites [17]. This perennial species has a self-compatibility mating system [18] and originated from a hybridization event between two diploid species: *A. lyrata* and *A. halleri* [19] [20]. The complete genome of the model plant *A. thaliana* [21], reference subgenomes of the diploid-progenitors *A. lyrata* subsp. *lyrata* [22] and *A. halleri* subsp. *gemmifera* [23], are available for use in analyzing *A. kamchatica*. Utilization of *A. kamcha-*

tica as an allopolyploid model species would unleash a treasure trove of genetic resources.

We applied Pool-Seq to the allopolyploid *A. kamchatica* by developing a bioinformatic pipeline from existing tools to simultaneously identify homeolog-specific and allele-specific polymorphisms in natural populations. Moreover, we assessed allele frequencies in each homeolog in the populations analyzed using the obtained Pool-Seq reads and compared these frequencies with individual-based estimates to validate the feasibility and utility of our protocol for cost-effective evaluation of nucleotide polymorphisms in allopolyploid populations.

2. Material and Methods

We applied a candidate gene approach to Pool-Seq by preparing Next Generation Sequencing (NGS) libraries consisting of pooled amplicons of eight target genes amplified from DNA samples representing 24 populations. Each sample was a pooled from multiple individuals possessing a population specific multiple identifier (MID) barcode, as detailed below.

2.1. NGS Library Preparation and Pool-Seq Procedure

We collected leaf tissue from 20 individuals of *A. kamchatica* subsp. *kamchatica* from each of 24 populations in central Honshu, Japan. Twenty-three of these populations were previously studied [16] (populations numbers 1, 2, 5, 8 - 21, and 23 - 29 in **Table 1** of their paper); the other population was Syomyodaki (latitude N36.5764˚, longitude E137.5186˚, altitude 1060 m). Genomic DNA from each individual was extracted from 15 mg of dried leaves using a DNeasy 96 Plant Kit or DNeasy Plant Mini Kit (Qiagen, Hilden, Germany), and concentration was then measured using a Quant-IT dsDNA HS Assay kit (Invitrogen, Life Technologies, Carlsbad, CA). Equal amounts (100 ng) of genomic DNA from each individual was pooled for each population, resulting in 2 μg (20 × 100 ng) of genomic DNA from each of the 24 populations being used as a template for amplification of the target genes.

Eight genes associated with flowering pathways or herbivore defense traits (*DFL*2, *GI*, *GL*1, *HEN*2, *MAM*1, *TTG*1, *CRY*1 and *PHYB*) were selected to be screened for nucleotide polymorphisms because these genes are considered to be single-copy genes, and putative divergent selective pressure associated with these traits in the target populations [16] could have resulted in nucleotide polymorphisms in the target genes. We used Primer3 [24] and conserved genome sequences between the *A. halleri* subsp. *gemmifera* [23] and *A. lyrata* subsp. *lyrata* [22] as a reference to design primers that would simultaneously amplify PCR products from both *A. halleri*-derived and *A. lyrata*-derived homeologs (hereafter *H*- and *L*-homeolog, respectively), with amplicon lengths of 390-590 bp. Primers consisted of gene specific sequence (**Table 1**) and 5'-appended M13 overhang (5'-CAGGGTTTTCCCAGTCACGAC-3') for forward primers, or 454

Table 1. Primer sequences designed for the amplification of eight target gene regions in the first round of PCR.

Primer set	Sequence (5'-3')	Target gene
CRY1_5_7	F: ATACAGTGTATTATAACATGATGGA R: TGAACTGGAGACGGCTTTCA	*CRY*1
DFL2_1_2	F: TTTTATTTCTTCATTGTAAATGGTCA R: TCCTCTGATTTCCCAGTAGATTACA	*DFL*2
GI_1_2	F: ATACAGTGTATTATAACATGATGGA R: TGAACTGGAGACGGCTTTCA	*GI*
GL1_	F: TTTTATTTCTTCATTGTAAATGGTCA R: TCCTCTGATTTCCCAGTAGATTACA	*GL*1
HEN2_3_4	F: ATACAGTGTATTATAACATGATGGA R: TGAACTGGAGACGGCTTTCA	*HEN*2
MAM1_1_2	F: TTTTATTTCTTCATTGTAAATGGTCA R: TCCTCTGATTTCCCAGTAGATTACA	*MAM*1
TTG1_1_3	F: ATACAGTGTATTATAACATGATGGA R: TGAACTGGAGACGGCTTTCA	*TTG*1
PHYB_01	F: TTTTATTTCTTCATTGTAAATGGTCA R: TCCTCTGATTTCCCAGTAGATTACA	*PHYB*

Adaptor B overhang (5'-CCTATCCCCTGTGTGCCTTGGCAGTCTCAG-3') for reverse primers. We performed PCR twice: the first for gene-specific amplification and the second for adapter incorporation. For the first round of PCR, we prepared 2 μl reaction volumes containing 0.2 ng of pooled genomic DNA from each population, previously dried on the bottom of the tubes [25], 200 μM dNTP mixture, 2.0 mM 1 × iProof HF Buffer (BIO-RAD, Hercules, CA), 0.04 U iProof High Fidelity DNA Polymerase (BIO-RAD, Hercules, CA), and 0.2 μM of the forward and reverse primers, with 6 μl of mineral oil overlaid. Thermocycling (with heated lid) was initiated with 98˚C for 30 sec, followed by 10 touchdown cycles of denaturation at 98˚C for 10 sec, annealing at 68˚C - 59˚C (decreasing by 1˚C per cycle) for 20 sec, and extension at 72˚C for 15 sec, followed by 35 equivalent cycles with annealing at 63˚C for 20 sec, and a final extension at 72˚C for 7 min. Two microliters of 10-times diluted PCR product from the first PCR was used as the template for the second PCR, which had a reaction volume of 10 μl and the same concentrations of reagents as the first PCR, except that we used a fusion primer consisting of 454 Adaptor A (5'-CCATCTCATCCCTGCGTGTCTCCGACTCAG-3') followed by a 10-base MID oligo and the above-mentioned M13 overhang as forward primer. The thermocycling conditions (with heated lid) were initiated at 98˚C for 30 sec, followed by 5 cycles of denaturation at 98˚C for 10 sec, annealing at 55˚C for 20 sec, and extension at 72˚C for 15 sec, followed by 35 cycles of denaturation at 98˚C for 10 sec and annealing and extension at 72˚C for 35 sec, and a final extension at 72˚C for 7 min. The PCR product from the second PCR for each population and target gene was purified with an Agencourt AMPure XP kit (Beck-

man Coulter, Milano, Italy), and then concentration was measured using a Quant-IT dsDNA BR Assay kit (Invitrogen, Life Technologies, Carlsbad, CA). Using the concentration data and the expected amplicon length, the molarity of the second amplicon was calculated as follows:

$$\text{Molecules}/\mu\text{l} = \frac{\text{sample concentration}\left[\text{ng}/\mu\text{l}\right] \times N_A}{656.6 \times 10^9 \times \text{expected amplicon length}\left[\text{bp}\right]}$$

where N_A is Avogadro's constant. We equalized the molarity of the 192 amplicons, which were amplified from each of 24 populations using primer sets for the eight target genes, and pooled all the amplicons to make a single library for a first sequencing. We followed the manufacturer's instructions for DNA quantity in the resulting library and sequencing procedures for the 454 GS Junior system (Roche, Basel, Switzerland).

The first 454-sequencing data for two genes, *CRY*1 and *PHYB*, had a relatively high rate of PCR chimeras (7.2%, referring to the output of the program UCHIME [26], as described below). For these two genes, we performed a second 454-sequencing run using a modified PCR protocol: no touch-down procedure, an extended elongation time to suppress incomplete primer extension [27], a lower number of cycles and a slower ramp speed [28]. Both the first and second PCR for the second sequencing run were performed in a total volume of 10 μl, containing 0.2 ng template DNA, 200 μM dNTP mixture, 1 × PrimeSTAR Buffer (TaKaRa Bio, Tokyo, Japan), and 0.3 μM of the primer pairs mentioned above. The first PCR was initiated with heated lid at 98˚C for 30 sec, followed by 30 cycles of denaturation at 98˚C for 10 sec, annealing at 52 or 59˚C (for *CRY*1 or *PHYB*, respectively) for 5 sec, and extension at 72˚C for 45 or 90 sec (for *CRY*1 or *PHYB*, respectively), followed by a final extension at 72˚C for 7 min. The second PCR (for both genes) was initiated with heated lid at 98˚C for 30 sec, followed by 5 cycles of denaturation at 98˚C for 10 sec, annealing at 56˚C for 5 sec, and extension at 72˚C for 45 sec, followed by 30 cycles of denaturation at 98˚C for 10 sec and extension at 72˚C for 50sec, followed by a final extension at 72˚C for 7 min. We equalized the molarity of the second amplicons, *i.e.* for each of *CRY*1 and *PHYB* among populations, and pooled them to make a 454 library for the second sequencing run, which was conducted in 454 GS Junior using approximately 10% of a plate following Gardner *et al.* [29]. For downstream analyses of *CRY*1 and *PHYB*, we only used reads from the second run, which had fewer PCR chimeras (under 0.1%).

2.2. Data Analysis Pipeline

The data processing workflow is shown in **Figure 1**. More details describing the data processing script are available in the **Appendices**. Procedures for analyzing genes with high sequence similarity have already been described, for example 16S rRNA analysis for assessing microbial diversity, which are based on use of assembler programs e.g. [30] [31]. However, we applied a mapping strategy to

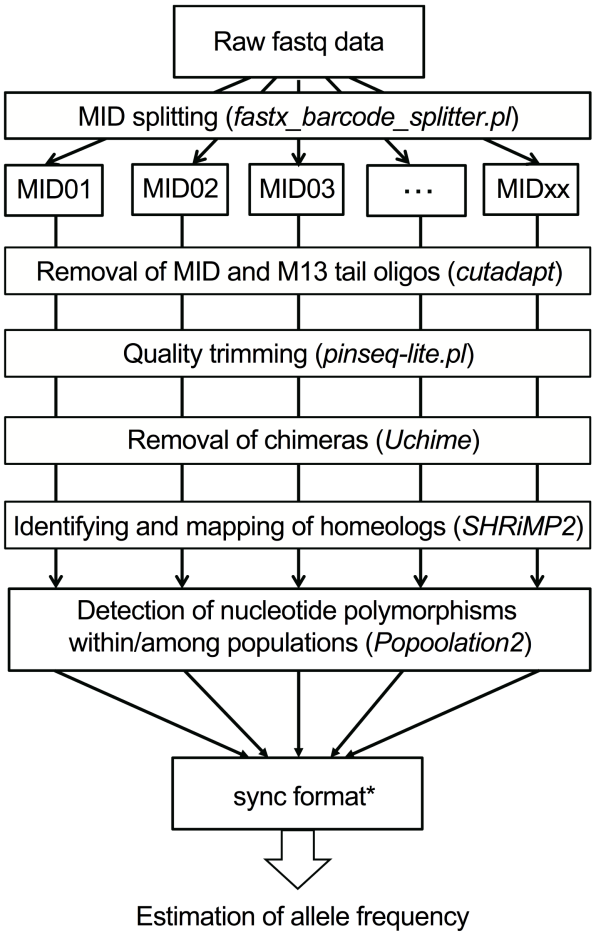

Figure 1. Workflow of our data analysis pipeline. *: sync format (sync format text file, see [12]).

identify homeologs, derived from the diploid-progenitors (*A. lyrata* and *A. halleri*), because existing subgenome information from these parental species was available. First, the sequencing reads were demultiplexed according to each specific MID-barcode (each MID barcode corresponds to a population) using the program FASTX-toolkit (available at http://hannonlab.cshl.edu/fastx_toolkit/links.html). The MID-barcodes and the M13 tail were trimmed from the reads using Cutadapt [32]. We then denoised all remaining reads (without a MID barcode and M13 tail) to avoid diversity overestimation caused by sequencing errors, including sequences with average quality score < Q27 or lengths shorter than 150 bp, using the program Prinseq-lite [33]. Potentially chimeric reads were subsequently eliminated using the program UCHIME [26]. The trimmed reads were assigned and mapped to each homeolog with the best mapping score between reads ("queries") and parental sequences ("references") using the mapping program SHRiMP2 [34]. Parental reference sequences for *H*- and *L*-homeologs of target genes (*i.e.* sequences predicted to be amplified by our primers) were obtained from genome sequences of *A. halleri* subsp. *gemmifera* [23] and *A. lyrata* subsp. *lyrata* [22], respectively.

Nucleotide polymorphisms were called with a Q20 variant quality score threshold using the mpileup command in the program SAMtools [35]. Information on all of the detected nucleotide polymorphisms across homeologs and populations was placed into a single file using the perl script mplileup2sync.pl in PoPoolation 2 [36], facilitating the estimation of allele frequencies for each homeolog and population. The effect of read length on the number of reads trimmed was assessed by generalized linear mixed model (GLMM) of the Poisson family, using the lmer function in the lme4 package in R version 3.3.1 [37], where genes, populations, and homeologs were set as random effects.

2.3. Validation of Pool-Seq

We evaluated the accuracy of estimates of allele frequencies derived from Pool-Seq for each population by conducting individual-based genotyping for two sites that were highly polymorphic both within and among populations. The first site was in the *GL1-H*-homeolog and had an A/C single nucleotide polymorphism (SNP). The second site was in the *PHYB-H*-homeolog and had 15-bp insertion-deletion polymorphism (indel). We genotyped the *GL1-H*-homeolog SNP for 20 individuals from each of four populations (Populations 12, 13 and 29 in Kenta *et al.* [16] and Syomyodaki described above) using SNP-SCALE [25] [38]. We genotyped the *PHYB-H*-homeolog indel for 20 individuals from each of five populations (Populations 12, 13, 16 and 20 in Kenta *et al.* [16] and Syomyodaki) using fragment analyses. We amplified the *A. halleri*-derived *PHYB* homeolog using target-specific PCR fusion primers (F: 5'-TGACTACGAATTTGATTTAGGCCT-3', R: 5'-U19-CTCTGGAGGCAGACCTTCAC-3') and FAM-labeled U19 universal primers, for 20 individuals from each of the five focal populations. PCR was conducted in 10 μl reaction mixtures containing approximately 1 ng of an individual's genomic DNA, 5 μl of 2 × AmpliTaq Gold 360 Master Mix (Applied Biosystems, Foster City, CA, USA), 0.5 μM of the forward primer, 0.3 μM of the reverse primer, and 0.5 μM of the fluorescent universal primer. The thermocycler was initiated with a first denaturation at 95˚C for 10 min, followed by 40 cycles of denaturation at 95˚C for 30 sec, annealing at 60˚C for 30 sec, and extension at 72˚C for 30 sec, followed by a final extension at 72˚C for 7 min. The indel alleles were determined by the size of the PCR products measured by capillary electrophoresis with an ABI PRISM 3100 sequencer (Applied Biosystems, Foster City, CA, USA).

3. Results and Discussion

We obtained a total of 25,011 reads from the target amplicons in the two sequencing runs. The sequencing reads were deposited at the DDBJ Sequence Read Archive under accession number DRA003062. Although a full plate was used in the first 454 GS Junior sequencing, it did not yield the expected number of reads (~100,000), because a large amount of byproduct reads with length less than 100

bp were obtained—possibly generated from a small amount of relatively large primer dimers that remained in the emulsion PCR (emPCR) template. A total of 21,047 reads were mapped to the references, giving an average mapping rate of 98.4% after quality trimming and removal of chimeras. The mean number of reads per population per homeolog was 61.4. There was substantial variation in the number of reads, ranging from 6 to 332, with a standard deviation of 53.7 among the 15 homeologs from the eight target genes (**Figure 2**), excluding the *GL*1-*L*-homeolog that had no reads because of primer mismatch. Such variation is likely attributable to our quantification method, using a DNA fluorometer rather than quantitative real-time PCR, to equalize PCR amplicons. We found that the number of reads decreased with an increase in the expected length of the PCR amplicon, even within pairwise homeologs ($P < 0.001$, **Figure 2**), suggesting basic difficulties underlying equalization of read numbers across different genes and homologs, because emPCR tends to capture shorter amplicons e.g. [39]. Kofler *et al.* (2016) points out, however, that variation in amplicon length of target genes with small-size indels had only a minor effect on the consistency of allele frequency estimates [40]. Thus, we assessed the Pool-Seq-based estimations of homeolog-specific allele frequencies, as described below. In the total target amplified length of 7125 bp for eight genes (*i.e.* pairwise homeologs except *GL*1-*L*-homeolog), we identified 144 putatively polymorphic sites which had at least 20 total reads including at least two minor-allele reads as threshold levels. Please note, however, that nucleotide polymorphisms detected in the two sequencing runs might be overestimated because filtering programs such as UCHIME cannot entirely remove chimeric DNA sequences. Of the 144 poly morphic sites, 70 sites were in the subgenome derived from *A. lyrata* and 74

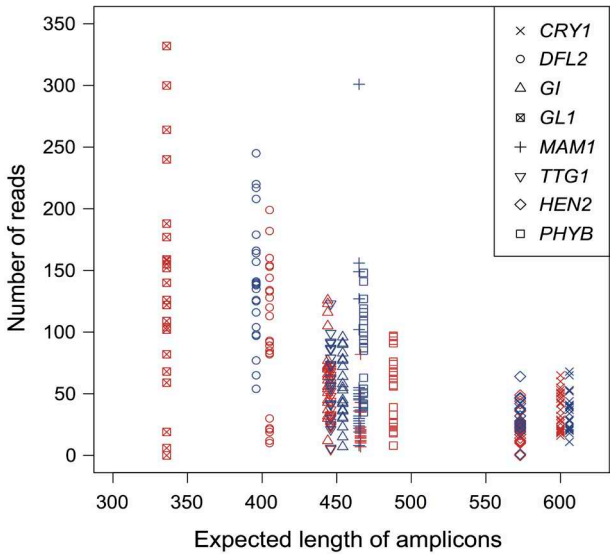

Figure 2. Number of reads per homeolog across 24 populations and the expected length of amplicons. Red points show *A. halleri*-derived homeologs and blue points show *A. lyrata*-derived homeologs.

were in the subgenome from *A. halleri*. The number of polymorphic sites was not significantly different between the homeologous genes ($P > 0.1$ by Wilcoxon signed rank test).

We found that the Pool-Seq estimations of homeolog-specific allele frequencies correlated well with those obtained by individual-based genotyping ($R^2 = 0.98$, $P < 0.001$, Figure 3). The maximum difference in allele frequency estimates between Pool-Seq and individual-based genotyping was 0.099. Segregating polymorphisms within a population, in which the most common allele had a frequency less than 0.9, were observed in three out of the eight tested populations (Figure 3), despite the self-compatibility of *A. kamchatica* [18]. Our results indicate the usefulness of Pool-Seq for estimating allele frequencies in allopolyploid populations, which will allow future applications of the technique in population genetics studies, such as detecting signatures of selection. As we detected the effect of read length on inconsistencies in read numbers between genes and between homeologs of a gene, we need to be careful when analyzing situations where there is a large difference in read length between alleles. This could cause inaccuracy in allele frequency estimates.

4. Conclusion

We developed laboratory and computational protocols to identify nucleotide polymorphisms within and among populations of allopolyploid species. We tested Pool-Seq as a method to simultaneously estimate allele frequencies for

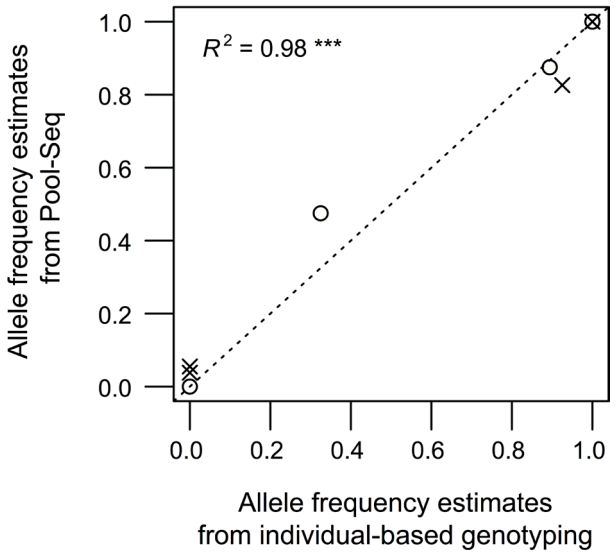

Figure 3. Allele frequency of two genes estimated by Pool-Seq and by individual-based genotyping in four populations of *A. kamchatica*. Open circles indicate the estimated frequency of SNP alleles (A/C) in the *A. halleri*-derived homeolog of *GL1*; cross marks indicate the estimated frequency of 15 bp indel alleles in the *A. halleri*-derived homeolog of *PHYB*. The dashed line shows a 1:1 relationship between Pool-Seq-based and individual-based estimations.

multiple populations, with no need for homeolog-specific PCR amplifications e.g. [20], using a bioinformatic pipeline designed to assign reads to homeologs originating from distinct diploid-progenitors. We showed that allele frequencies estimated by Pool-Seq correlated well with precise allele frequencies determined by individual genotyping and without any systematic biases. Of particular note was that chimera formation in the PCR process is a potential problem for precise estimation of allele frequencies, especially in allopolyploids, because simultaneous PCR amplification of pairwise homeologs with high sequence similarity could potentially generate more PCR chimeras in an allopolyploid compared to a diploid. We have shown these problems can be mitigated by using an optimized PCR protocol before computational filtering of chimera reads. Although GS Junior and other 454 platforms (Roche) have not been supported by the manufacturer since 2016, our methodology using the 454 platform is applicable to other NGS platforms including the Ion PGM (Thermo Fisher), which is completely compatible with 454 libraries, and Miseq (Illumina). With allele frequency estimates being key to population genetic analysis such as detecting signatures of selection, Pool-Seq provides a cost-effective approach for identifying nucleotide polymorphisms among a large number of individuals and genes, particularly in allopolyploid species.

Acknowledgements

We would like to thank H. Yuki, K. Hatada, and M. Suzuki for DNA experiments; I. Sakaguchi for collecting samples; and to Dr. A. Kurokawa and Dr. H. Mori for data analysis advice. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics. This work was supported by JSPS (23770016 to A.S.H.), Research and Education Funding for Japanese Alps Inter-Universities Cooperative Project, MEXT Japan KAKENHI (16H06469, 16H06464, 16K21727 to J.S. and K.K.S., 26113709 to K.K.S., 22770231 to T.K. and J.S., and 221S0002), HFSP to J.S. and K.K.S., and Swiss National Science Foundation and URPP Evolution in Action to R.S.I. and to K.K.S., and JST CREST (number JPMJCR16O3) to K.K.S., J.S. and T.K.

Conflict of Interest Declaration

The authors declare that they have no conflict of interest regarding the publication of this paper.

References

[1] Bullini, L. (1994) Origin and Evolution of Animal Hybrid Species. *Trends in Ecology & Evolution*, **9**, 422-426. https://doi.org/10.1016/0169-5347(94)90124-4

[2] Ramsey, J. and Schemske, D.W. (1998) Pathways, Mechanisms, and Rates of Polyploid Formation in Flowering Plants. *Annual Review of Ecology and Systematics*, **29**, 467-501. https://doi.org/10.1146/annurev.ecolsys.29.1.467

[3] Soltis, D.E., Visger, C.J. and Soltis, P.S. (2014) The Polyploidy Revolution Then and Now: Stebbins Revisited. *American Journal of Botany*, **101**, 1057-1078.

https://doi.org/10.3732/ajb.1400178

[4] Stebbins, G. (1984) Polyploidy and the Distribution of the Arctic-Alpine Flora: New Evidence and a New Approach. *Botanica Helvetica*, **94**, 1-13.

[5] International Wheat Genome Sequencing Consortium (2014) A Chromosome-Based Draft Sequence of the Hexaploid Bread Wheat (*Triticum aestivum*) Genome. *Science*, **345**, 286. https://doi.org/10.1126/science.1251788

[6] Buggs, R.J., Chamala, S., Wu, W., Gao, L., May, G.D., Schnable, P.S., Soltis, D.E., Soltis, P.S. and Barbazuk, W.B. (2010) Characterization of Duplicate Gene Evolution in the Recent Natural Allopolyploid *Tragopogon miscellus* by Next-Generation Sequencing and Sequenom iPLEX MassARRAY Genotyping. *Molecular Ecology*, **19**, 132-146. https://doi.org/10.1111/j.1365-294X.2009.04469.x

[7] Session, A.M., Uno, Y., Kwon, T., Chapman, J.A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., van Heeringen, S.J., Quigley, I., Heinz, S., Ogino, H., Ochi, H., Hellsten, U., Lyons, J.B., Simakov, O., Putnam, N., Stites, J., Kuroki, Y., Tanaka, T., Michiue, T., Watanabe, M., Bogdanovic, O., Lister, R., Georgiou, G., Paranjpe, S.S., van Kruijsbergen, I., Shu, S., Carlson, J., Kinoshita, T., Ohta, Y., Mawaribuchi, S., Jenkins, J., Grimwood, J., Schmutz, J., Mitros, T., Mozaffari, S.V., Suzuki, Y., Haramoto, Y., Yamamoto, T.S., Takagi, C., Heald, R., Miller, K., Haudenschild, C., Kitzman, J., Nakayama, T., Izutsu, Y., Robert, J., Fortriede, J., Burns, K., Lotay, V., Karimi, K., Yasuoka, Y., Dichmann, D.S., Flajnik, M.F., Houston, D.W., Shendure, J., DuPasquier, L., Vize, P.D., Zorn, A.M., Ito, M., Marcotte, E.M., Wallingford, J.B., Ito, Y., Asashima, M., Ueno, N., Matsuda, Y., Veenstra, G.J., Fujiyama, A., Harland, R.M., Taira, M. and Rokhsar, D.S. (2016) Genome Evolution in the Allotetraploid Frog *Xenopus laevis*. *Nature*, **538**, 336-343. https://doi.org/10.1038/nature19840

[8] Harper, A.L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., Hattori, C., Werner, P. and Bancroft, I. (2012) Associative Transcriptomics of Traits in the Polyploid Crop Species Brassica Napus. *Nature Biotechnology*, **30**, 798-802. https://doi.org/10.1038/nbt.2302

[9] Akama, S., Shimizu-Inatsugi, R., Shimizu, K.K. and Sese, J. (2014) Genome-Wide Quantification of Homeolog Expression Ratio Revealed Nonstochastic Gene Regulation in Synthetic Allopolyploid *Arabidopsis*. *Nucleic Acids Research*, **42**, e46. https://doi.org/10.1093/nar/gkt1376

[10] Buggs, R.J., Renny-Byfield, S., Chester, M., Jordon-Thaden, I.E., Viccini, L.F., Chamala, S., Leitch, A.R., Schnable, P.S., Barbazuk, W.B., Soltis, P.S. and Soltis, D.E. (2012) Next-Generation Sequencing and Genome Evolution in Allopolyploids. *American Journal of Botany*, **99**, 372-382. https://doi.org/10.3732/ajb.1100395

[11] Dufresne, F., Stift, M., Vergilino, R. and Mable, B.K. (2014) Recent Progress and Challenges in Population Genetics of Polyploid Organisms: An Overview of Current State of the Art Molecular and Statistical Tools. *Molecular Ecology*, **23**, 40-69. https://doi.org/10.1111/mec.12581

[12] Kofler, R., Orozco-Terwengel, P., De Maio, N., Pandey, R., Nolte, V., Futschik, A., Kosiol, C., Schlötterer, C. and Kayser, M. (2011) PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLoS One*, **6**, 587-591. https://doi.org/10.1371/journal.pone.0015925

[13] Bastide, H., Betancourt, A., Nolte, V., Tobler, R., Stöbe, P., Futschik, A. and Schlötterer, C. (2013) A Genome-Wide, Fine-Scale Map of Natural Pigmentation Variation in *Drosophila melanogaster*. *PLoS Genetics*, **9**, e1003534. https://doi.org/10.1371/journal.pgen.1003534

[14] Turner, T.L., Bourne, E.C., Von Wettberg, E.J., Hu, T.T. and Nuzhdin, S.V. (2010) Population Resequencing Reveals Local Adaptation of Arabidopsis Lyrata to Ser-

pentine Soils. *Nature Genetics*, **42**, 260-263. https://doi.org/10.1038/ng.515

[15] Khan, A., Belfield, E.J., Harberd, N.P. and Mithani, A. (2016) HANDS2: Accurate Assignment of Homoeallelic Base-Identity in Allopolyploids despite Missing Data. *Scientific Reports*, **6**, 29234. https://doi.org/10.1038/srep29234

[16] Kenta, T., Yamada, A. and Onda, Y. (2011) Clinal Variation in Flowering Time and Vernalisation Requirement across a 3000-m Altitudinal Range in Perennial *Arabidopsis kamchatica* ssp. *kamchatica* and Annual Lowland Subspecies *kawasakiana*. *Journal of Ecosystem* & *Ecography*, **S6**, 1.

[17] Sugisaka, J. and Kudoh, H. (2007) Breeding System of the Annual Cruciferae, *Arabidopsis kamchatica* subsp. *kawasakiana*. *Journal of Plant Research*, **121**, 65-68. https://doi.org/10.1007/s10265-007-0119-7

[18] Tsuchimatsu, T., Kaiser, P., Yew, C.L., Bachelier, J.B. and Shimizu, K.K. (2012) Recent Loss of Self-Incompatibility by Degradation of the Male Component in Allotetraploid *Arabidopsis kamchatica*. *PLoS Genetics*, **8**, e1002838. https://doi.org/10.1371/journal.pgen.1002838

[19] Shimizu, K.K., Fujii, S., Marhold, K., Watanabe, K. and Kudoh, H. (2005) *Arabidopsis kamchatica* (Fisch. ex DC.) K. Shimizu & Kudoh and *A. kamchatica* subsp. *kawasakiana* (Makino) K. Shimizu & Kudoh, New Combinations. *Acta Phytotaxonomica et Geobotanica*, **56**, 163-172.

[20] Shimizu-Inatsugi, R., Lihova, J., Iwanaga, H., Kudoh, H., Marhold, K., Savolainen, O., Watanabe, K., Yakubov, V.V. and Shimizu, K.K. (2009) The Allopolyploid *Arabidopsis kamchatica* Originated from Multiple Individuals of *Arabidopsis lyrata* and *Arabidopsis halleri*. *Molecular Ecology*, **18**, 4024-4048. https://doi.org/10.1111/j.1365-294X.2009.04329.x

[21] Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2012) The *Arabidopsis* Information Resource (TAIR): Improved Gene Annotation and New Tools. *Nucleic Acids Research*, **40**, D1202-D1210. https://doi.org/10.1093/nar/gkr1090

[22] Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D. and Guo, Y.L. (2011) The *Arabidopsis lyrata* Genome Sequence and the Basis of Rapid Genome Size Change. *Nature Genetics*, **43**, 476-481. https://doi.org/10.1038/ng.807

[23] Briskine, R.V., Paape, T., Shimizu-Inatsugi, R., Nishiyama, T., Akama, S., Sese, J. and Shimizu, K.K. (2016) Genome Assembly and Annotation of *Arabidopsis halleri*, a Model for Heavy Metal Hyperaccumulation and Evolutionary Ecology. *Molecular Ecology Resources*. https://www.readbyqxmd.com/read/27671113/genome-assembly-and-annotation-of-arabidopsis-halleri-a-model-for-heavy-metal-hyperaccumulation-and-evolutionary-ecology

[24] Rozen, S. and Skaletsky, H. (1999) Primer3 on the WWW for General Users and for Biologist Programmers. In: Misener, S. and Krawetz, S.A., Eds., *Bioinformatics Methods and Protocols*, Springer, Berlin, 365-386. https://doi.org/10.1385/1-59259-192-2:365

[25] Kenta, T., Gratten, J., Haigh, N.S., Hinten, G.N., Slate, J., Butlin, R.K. and Burke, T. (2008) Multiplex SNP-SCALE: A Cost-Effective Medium-Throughput Single Nucleotide Polymorphism Genotyping Method. *Molecular Ecology Resources*, **8**, 1230-

1238. https://doi.org/10.1111/j.1755-0998.2008.02190.x

[26] Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. and Knight, R. (2011) UCHIME Improves Sensitivity and Speed of Chimera Detection. *Bioinformatics*, **27**, 2194-2200. https://doi.org/10.1093/bioinformatics/btr381

[27] Kanagawa, T. (2003) Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR). *Journal of Bioscience and Bioengineering*, **96**, 317-323. https://doi.org/10.1016/S1389-1723(03)90130-7

[28] Stevens, J.L., Jackson, R.L. and Olson, J.B. (2013) Slowing PCR Ramp Speed Reduces Chimera Formation from Environmental Samples. *Journal of Microbiological Methods*, **93**, 203-205. https://doi.org/10.1016/j.mimet.2013.03.013

[29] Gardner, M.G., Fitch, A.J., Bertozzi, T. and Lowe, A.J. (2011) Rise of the Machines-Recommendations for Ecologists When Using Next Generation Sequencing for Microsatellite Development. *Molecular Ecology Resources*, **11**, 1093-1101. https://doi.org/10.1111/j.1755-0998.2011.03037.x

[30] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q. and Liu, Y. (2012) SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read de Novo Assembler. *Gigascience*, **1**, 18. https://doi.org/10.1186/2047-217X-1-18

[31] Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y. (2012) Meta Velvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads. *Nucleic Acids Research*, **40**, e155. https://doi.org/10.1093/nar/gks678

[32] Martin, M. (2011) Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. *EMBnet Journal*, **17**, 10-12. https://doi.org/10.14806/ej.17.1.200

[33] Schmieder, R. and Edwards, R. (2011) Quality Control and Preprocessing of Metagenomic Datasets. *Bioinformatics*, **27**, 863-864. https://doi.org/10.1093/bioinformatics/btr026

[34] Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: Accurate Mapping of Short Color-Space Reads. *PLoS Computational Biology*, **5**, e1000386. https://doi.org/10.1371/journal.pcbi.1000386

[35] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map Format and SAMtools. *Bioinformatics*, **25**, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

[36] Kofler, R., Pandey, R.V. and Schlotterer, C. (2011) PoPoolation2: Identifying Differentiation between Populations Using Sequencing of Pooled DNA Samples (Pool-Seq). *Bioinformatics*, **27**, 3435-3436. https://doi.org/10.1093/bioinformatics/btr589

[37] R Core Team (2016) R: A Language and Environment for Statistical Computing. R Core Team, R Foundation for Statistical Computing, Vienna, Austria.

[38] Hinten, G.N., Hale, M.C., Gratten, J., Mossman, J.A., Lowder, B.V., Mann, M.K. and Slate, J. (2007) SNP-SCALE: SNP Scoring by Colour and Length Exclusion. *Molecular Ecology Notes*, **7**, 377-388. https://doi.org/10.1111/j.1471-8286.2006.01648.x

[39] Bybee, S.M., Bracken-Grissom, H., Haynes, B.D., Hermansen, R.A., Byers, R.L., Clement, M.J., Udall, J.A., Wilcox, E.R. and Crandall, K.A. (2011) Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. *Genome Biology and Evolution*, **3**, 1312-1323. https://doi.org/10.1093/gbe/evr106

[40] Kofler, R., Nolte, V. and Schlotterer, C. (2016) The Impact of Library Preparation Protocols on the Consistency of Allele Frequency Estimates in Pool-Seq Data. *Molecular Ecology Resources*, **16**, 118-122. https://doi.org/10.1111/1755-0998.12432

Appendixes

Appendix 1. An Example of Shell-Script for Bioinformatics Pipeline

```
# Analyzing_PoolSeq.sh
# This file is an example of shell script to detect nucleotide polymorphisms from the Pool-seq data (DRA003062).
# This script is free software: you can redistribute it and/or modify it. # under the terms of the GNU Lesser General Public License
# as published by the Free Software Foundation, either version 3 of
# the License, or (at your option) any later version.
# This script is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of # MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
#Requirements (external programs)
# • Perl: It is likely that Perl is already installed.
# http://www.perl.org/get.html
# • fastx_barcode_splitter.pl in FASTX-Toolkit: A perl script for splitting MID barcodes.
# http://hannonlab.cshl.edu/fastx_toolkit/
# • Cutadapt: A command tool for removing adapter sequences.
# https://cutadapt.readthedocs.org/en/stable/
# • SAMtools: A suite of programs for interacting with high-throughput sequencing data.
# http://samtools.sourceforge.net
# • SHRiMP2: Read mapping program.
# http://compbio.cs.toronto.edu/shrimp/
# • UCHIME: Tools for detecting chimeric sequence
# http://drive5.com/usearch/manual/uchime_algo.html):
# • prinseq-lite.pl: A perl script that can be used to filter, reformat, or trim sequence data.
# http://prinseq.sourceforge.net
# • Popoolation2: A software specifically designed for the comparison of populations with Pool-Seq data.
# https://sourceforge.net/p/popoolation2/wiki/Main/
# Set full path to Cutadapt, SAMtools, SHRiMP2, and UCHIME executables (cutadapt, samtools, gmapper, and usearch, respectively).
# Set full path to the perl scripts (fastx_barcode_splitter.pl and prinseq-lite.pl).
# Set full path to the software Popoolation2.
#To run this script SCRIPT (in a Linux environment):
# Copy this script into a new directory along with:
# 1) A fastq file of the Pool-Seq data, named "poolseq.fastq"
# 2) A text file of MID barcodes, names "MID.txt" (see Appendix S2 for example)
```

```
# 3) A fasta file of reference sequences of target genes, names “refer
ence_genes.fasta” (see Appendix 3 for example) #You may need to make this
script and the other two programs executable, or able
#to be recognized as programs. To do this, run the following command:
#
# chmod +x Analyzing_PoolSeq.sh #
# Finally, to run the script, type the following command:
# ./Analyzing_PoolSeq.sh
#!/bin/bash
#definition for a fasta file of reference sequences
ref=reference_genes
#To define M13 universal sequence
M13_seq=CAGGGTTTTCCCAGTCACGAC
#To prepare a list of bam files for mpileup
target_all=" "
target2="/"
target3=".bam "
mkdir -p MID_all
#To split MID-barcodes using FASTX toolkit
cat poolseq.fastq | fastx_barcode_ splitter.pl --bcfile MID_list.txt --bol
--mismatches 2 --prefix MID_all/ --suffix ".fastq"
#To process into MID-specific reads
while read MID_no MID_seq; do
target_all=$target_all$MID_no$target2$MID_no$target3
if [ $MID_no = "#MID"] then
echo $MID_no $MID_seq else
echo $MID_no $MID_seq
#To remove MID barcode
cutadapt -g $MID_seq MID_all/$MID_no.fastq >
MID_all/$MID_no.noMID.fastq
mkdir-p $MID_no
mv MID_all/$MID_no.fastq $MID_no/
mv MID_all/$MID_no.noMID.fastq $MID_no/
cp $ref.fas $MID_no/
cd $MID_no
#To remove M13 universal sequence
cutadapt -g $M13_seq $MID_no.noMID.fastq > $MID_no.noMID_M13.fastq
prinseq-lite.pl -fastq $MID_no.noMID_M13.fastq -min_len 150
-trim_qual_right 27 -trim_left 33 -out_format 4 -out_good $MID_no.trim
-out_bad null
#To remove chimeric sequence
usearch -uchime $MID_no.trim.fasta --db $ref.fas --nonchimeras
$MID_no.nonchimera.fasta --log uchime.log
```

```
#To assign and map reads to homeologs: gmapper included in SHRiMP2
gmapper $MID_no.nonchimera.fasta $ref.fas -r 454-E > $MID_no.sam
samtools view -q 20 -bS $MID_no.sam | samtools sort - $MID_no
samtools index $MID_no.bam
echo
cd ../ fi
done < <(tail-n +3 MID_list.txt) #list of MID sequence
rm -rf MID_all
#To write information of all found nucleotide polymorphisms in a single file "AmongPops.sync"
echo $target_all
samtools mpileup-B $target_all > AmongPops.mpileup
#mpileup2sync.pl is included in Popoolation2
#You must set full path to the software Popoolation2
perl mpileup2sync.pl--fastq-type sanger--min-qual 20 --input Among
Pops.mpileup-- output AmongPops.sync
#See Kofler et al. (2011) for the file format of *.sync
```

Appendix 2. A List of MID Barcodes

MID_05 ATCAGACACG
MID_06 ATATCGCGAG
MID_07 CGTGTCTCTA
MID_08 CTCGCGTGTC
MID_09 TAGTATCAGC
MID_10 TCTCTATGCG
MID_11 TGATACGTCT
MID_12 TACTGAGCTA
MID_13 CATAGTAGTG
MID_14 CGAGAGATAC
MID_15 ATACGACGTA
MID_16 TCACGTACTA
MID_17 CGTCTAGTAC
MID_18 TCTACGTAGC
MID_19 TGTACTACTC
MID_20 ACGACTACAG
MID_21 CGTAGACTAG
MID_22 TACGAGTATG
MID_23 TACTCTCGTG
MID_25 TCGTCGCTCG
MID_26 ACATACGCGT
MID_27 ACGCGAGTAT
MID_28 ACTACTATGT
MID_30 AGACTATACT